Secure and Scalable Systems for Research Data Management

Paul Walsh, Senior Visiting Research Fellow

August 2014

Outline

- Context of Research
- Case Study
  - Neonatal sepsis
- Solution Four S's
  - Speed, Security, Scalability, Simplicity
- Conclusions

Context- Neonatal Infection

- **Sepsis** immune system's response to a serious infection

- 37,000 deaths a year in the UK and millions of deaths globally
- Without new and effective methods the problem will escalate

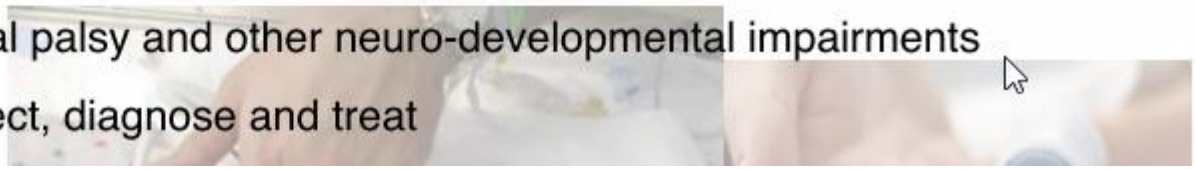
Context- Neonatal Infection

- **ClouDx-i** is an EU FP7 research project concerned with developing host pathogen response biomarkers for neonatal infection
- NSilico is tasked with Data Management
- See www.cloudxi.eu

ClouDx-i

Context- Neonatal Infection

- Infections (mainly bacterial) during the fragile neonatal period (0-10 days)
  - Group B streptococcus
  - E. coli and other gram -ve bacteria
    - Coagulase -ve staphylococcus
- Due to deficiencies in both innate and adaptive immunity
- Very common in preterm and low birth weight infants
- Infection remains an important source of morbidity and mortality in neonates and infants, accounting for more than half of all deaths worldwide of children younger than age 5
- Can lead to cerebral palsy and other neuro-developmental impairments
- Very difficult to detect, diagnose and treat



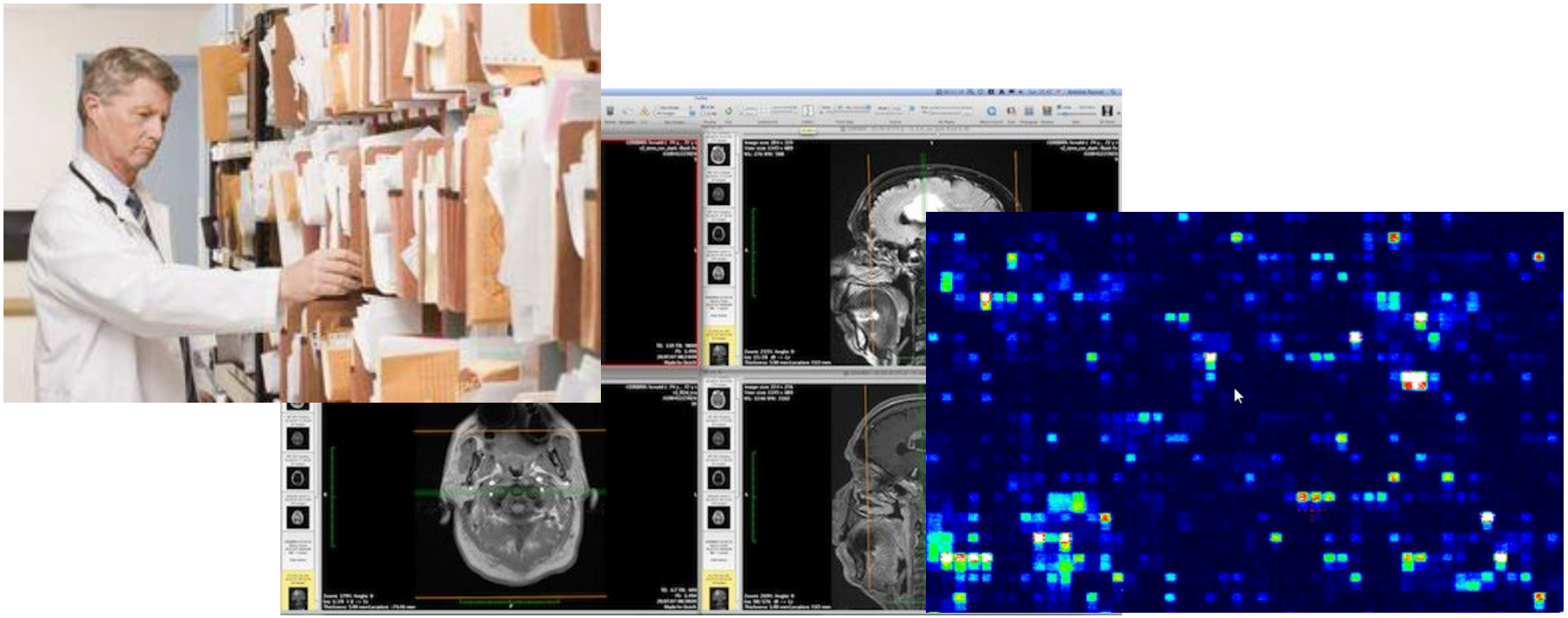
Division of
PATHWAY MEDICINE

Context- Neonatal Infection

- Can a few drops of blood provide insight into the pathogenesis of infection?
- Applying systems analysis can we derive a host-signal of infection?
- What can we learn about neonatal immunity?
- Can this improve the predictive power for diagnosis?
- Can we glean any therapeutic insights?

Context- Heterogeneous Big Data

- Numerous big data types that must be handled securely….



Clinical Health Records, DICOM, Genomic, Host DNA/RNA and pathogen.

Large Data Sets Need to be Managed

Sample collection & experimental design		*from blood samples (easy to collect) to brain tissue (hard to collect)*	~$100 onwards		from a few hours to several days
Sequencing		*library preparation + running the sequencer (whole dual flow cell)*	~$6500 = ~$500 + ~$6000	~380M reads/lane; 1 individual: ~1140M total reads (~3 lanes for a 30x coverage); ~250Gb (intermediate files)	~11-12 day
Data reduction & management	Data storage, low-level processing	*Alignment (transfer* and storing raw data + mapping)*	~$40 = ~$33 + ~$7	300Gb (BAM file)	~1/2 day *** (including transferring 250Gb FASTQ ~7.5 hrs)
		(data transfer and storage for 10 days); ***	~$40		~8.5 hrs
	High-level summaries***	*SNP calling (compute + transfer out)*	<$5 = ~$4 + ~$0.60	< 1Gb	~3 hrs
		Indel calling (compute + transfer out)	<$35 = ~$32 + ~$0.60	< 1Gb	~1 day
		SV calling (compute + transfer out)	<$35 = ~$32 + ~$0.60	< 1Gb	~1 day
Downstream analyses			>$100K	~310Gb	months

Large Data Sets Need to be Managed

Sample collection & experimental design		*from blood samples (easy to collect) to brain tissue (hard to collect)*	~$100 onwards		from a few hours to several days
Sequencing		*library preparation + running the sequencer (whole dual flow cell)*	~$6500 = ~$500 + ~$6000	~380M reads/lane; 1 individual: ~1140M total reads (~3 lanes for a 30x coverage); ~250Gb (intermediate files)	~11-12 day
Data redu… & …ment	…orage, low-processing	*Alignment (transfe… and sto…ing raw data m…ing)*	~$40 = ~$33 + ~$7	300Gb (BAM file)	~1/2 day *** (including transferring 250Gb FASTQ ~7.5 hrs) ~8.5 hrs
	…sum…		…$4 + ~$0.60	< 1Gb	~3 hrs
			…+ ~$0.60	< 1Gb	~1 day
		S…ling (co…te + tra…er out)	<$35 =	< 1Gb	~1 day
Downstream analyses			>$100K	~310Gb	months

Storage for single experiment is over 300Gb

Large Data Sets Need to be Managed

- Just to archive the sequences from a single run, the order of 300 GB (1-2 copies of the basic gzipped FASTQ data).
- For 10 data sets each month would require about 3 TB of permanent archival space.
- We aim to address thousands of cases.

http://ivory.idyll.org/blog/how-much-compute-ngs.html

Solution – Cloud Computing

- NSilico is the provides cloud based data management and analytics software for the lifesciences and healthcare industries.
- Core values:
  - ***S**implicity*
  - ***S**peed*
  - ***S**calability*
  - ***S**ecurity*

- Two core product areas:
  - Bioinformatics
  - Patient Record Management

- We are currently working with DPM to leverage these technologies for neonatal care within an FP7 ClouDx-i

Bioinformatics: *Simplicity™*

- Integrated with Big Data cloud storage AWS S3, DB, BaseSpace
- Easy analysis of raw "omic" data Enables non-specialists to easily manage bioinformatics workflows and data

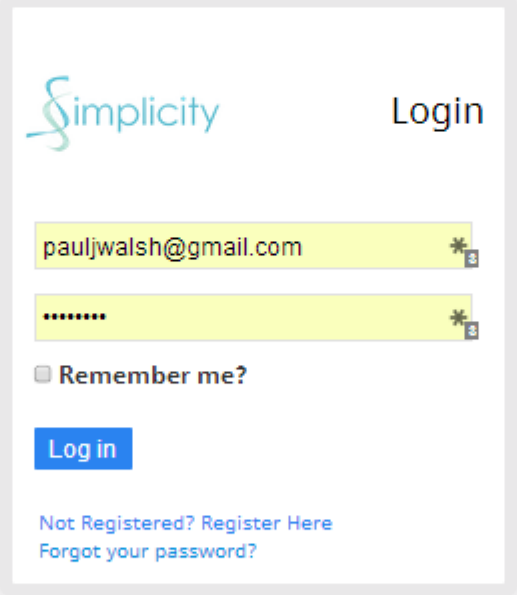
Bioinformatics: *Simplicity™*

- Massively reduces research cycle-times..
- Complete Audit Trail on all data.
- Short demo of DPM/RIE sample report ……..see http://youtu.be/8oyDu07zrKU

Welcome to *Simplicity*™ v1.0

Simplicity™ is the world's most powerful, yet easy-to-use bioinformatics tool which enables non-specialists to reliably run bioinformatics workflows in a simple, safe, speedy and secure manner.

Specifically, *Simplicity*™ v1.0 enables automatic analysis, annotation and visualisation of prokaryote data. It is aimed at microbiologists who lack an in-depth knowledge of bioinformatics. *Simplicity*™ v1.0 enables them to generate comprehensive reports from their raw data with just a few clicks.



Please choose the type of file you wish to process:

Assembled sequences in FASTA file format

Unassembled sequences in Illumina file format

GO

Simplicity

Prokaryote Workflow

Name of analysis: Analysis of Neonatal Sample X123

Abstract:

Analysis of Neonatal Sample X123

Upload sequence data

Dropbox Other File Sharing System

Or else select a library from the select box

None

Max allowed error Rate (%): 5

Minimum lenght of trimmed read: 0

Assembly Options (Spades):

Use post assembly tool MismatchCorrector to try reduce mistakes and short redels: (•) yes ( ) no

Reads file type: Files with left and right read library

Genetic code:

(•) The Bacterial, Archaeal and Plant Plastid Code (translation table=11)

( ) The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (translation table=4)

Go with default settings or expand to configure

Trial users will be limited to 500 ORF's for Protein similarity search with Gene Ontology

- [x] Gene prediction (Glimmer)
- [x] Protein similarity search with Gene Ontology
- [x] Multiple alignment (Clusted Omega)
- [x] Nucleotide similarity search with NCBI nt
- [x] Cath search
- [ ] pFam
- [x] Phylogenetic trees (Phangorn)

GO

Adapter	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
AdapterRev	
Quality cut off	10
Max allowed error Rate	5
Minimum length of trimmed read	0

Spades

Cell type	Single Cell
Reduced mistakes and short indels	yes
Reads file type	Files with left and right read library

Tools and Results:

Tool	Tool Number	Results	Settings	Status
FastQC	1			
Cutadapt	2			
FastQC after trimming	3			
Spades	4			
Quast	5			
Glimmer	6			
Cusp	7			
Gview	8			
Blastn NCBI nt	9			
Blastp with Gene Ontology	10			
Cath Domain Search	11			
pFam	12			
ClustalO	13			
Phangorn R	14			

Adapter ACACTCTTTCCCTACACGACGCTCTTCCGATCT

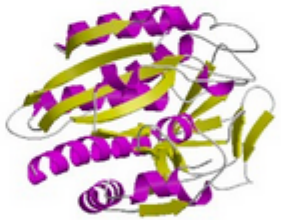
Cath Domain Search

Expert caption | All Blasted ORFs | Filter by Gene Ontology terms:

637 results

>6_[4350 - 5588]
>8_[6917 - 7795]
>11_[10429 - 11904]
>12_[12348 - 13997]
>15_[15663 - 16688]
>16_[16698 - 18890]
>17_[18991 - 20799]
>18_[20800 - 20970]
>19_[20993 - 22294]
>20_[22304 - 24115]
>21_[24108 - 25766]
>23_[26984 - 27202]
>24_[27297 - 27737]
>25_[27734 - 28168]
>36_[35633 - 38044]
>38_[39445 - 40104]
>39_[40134 - 41366]
>42_[43573 - 44484]
>43_[44577 - 45860]
>44_[45929 - 47032]
>45_[47144 - 48232]
>46_[48298 - 48897]
>47_[48933 - 49682]

Selected ORF Input

Accession_No	% ID	E-value	Homologous	Cath Code	Image
1vixB01	58	4.67267e-87			
1vixA01	58	5.6948e-87			

```
Class  =    | Archetype  =    | Topology  =
Score = 666   Bits (288.976),    Expect (e-value)  4.67267e-87   Identities = 122    Positives = 159
Query      9      ELLERFLHYVSFHTQSKPHAKHSPSSVGQMKLAMQLQKELIQLGLENVEVSKYAVVTAFLPANDPNLTKTIGLVAHLDTSPQCSGK
                  +LLERFL+YVS TQSK + PS+ KL L+++L ++GL NV +S+ + A LPAN P IG ++H+DTSP CSGKNV P+++E YRG
Subject    5      KLLERFLNYVSLDTQSKAGVRQVPSTEGQWKLLHLLKEQLEEMGLINVTLSEKGTLMATLPANVPGDIPAIGFISHVDTSPDCSGK

Class  =    | Archetype  =    | Topology  =
```

Adapter ACACTCTTTCCCTACACGACGCTCTTCCGATCT

AdapterRev

Quality cut off 10

Settings

Nucleotide Similarity Search (blastn) of NCBI nt database

Finds regions of similarity between biological sequences by executing BLAST with program, database, a query sequence and options.

Select protein databases

☑ NCBI nt (nucleotide sequence database, with entries from all traditional divisions of GenBank, EMBL, and DDBJ)

Set your parameters

Match/Mismatch 1,-3 ▼ Gap Open 11 ▼ Gap Extend 2 ▼ Expectation Threshold 1e-1 ▼ Alignments 10 ▼

Reports

Simplicity Report - PDF-XChange Viewer

File Edit View Document Comments Tools Window Help

Simplicity Report | 2013_GenomeMed_Miller_Metagenomics_Pu... | 57302053 | 57302053 (2) | 57164500 | 57302053 (3) | ECDU-capabilities-200914 | 1-s2 0-S0966842X14000183-main (2)

Bioinformatics Analysis of Staphylococcus Aureus

Paul Walsh[1], *Brian Kelly*[1]

Abstract

This presents the results of a bioinformatics analysis of Staphylococcus aureus is a widely distributed human pathogen capable of infecting almost every ecological niche of the host. As a result, it is responsible for causing many different diseases. These results were generated by the Simplicity™ software service, which cleaned, validated and assembled the reads. It then performed gene prediction, protein structure classification, protein domain classification, multiple alignment and local alignment and linked these results with the Gene Ontology database (GO).

Date started: 02/Jul/2014 17:19 **Date finished:** 03/Jul/2014 11:54

[1]*NSilico Life Science Ltd, Rubicon Centre, Bishopstown, Cork, Ireland, paul.walsh@nsilico.com*

Methods & findings

The results presented in this report were produced by the tools listed below. The tools and report generation were managed by *Simplicity™[1].*

Reads Quality

FastQC [2] is a high throughput quality control tool for sequence data and gives a quick impression of the data.

Reads Library type: Paired end
File type: Conventional base calls
Encoding: Sanger / Illumina 1.9

Filename	sol5675_s3_l001_r1_001.fastq.gz	sol5675_s3_l001_r2_001.fastq.gz
Total Sequences	1471060	1471060
Filtered Sequences	0	0

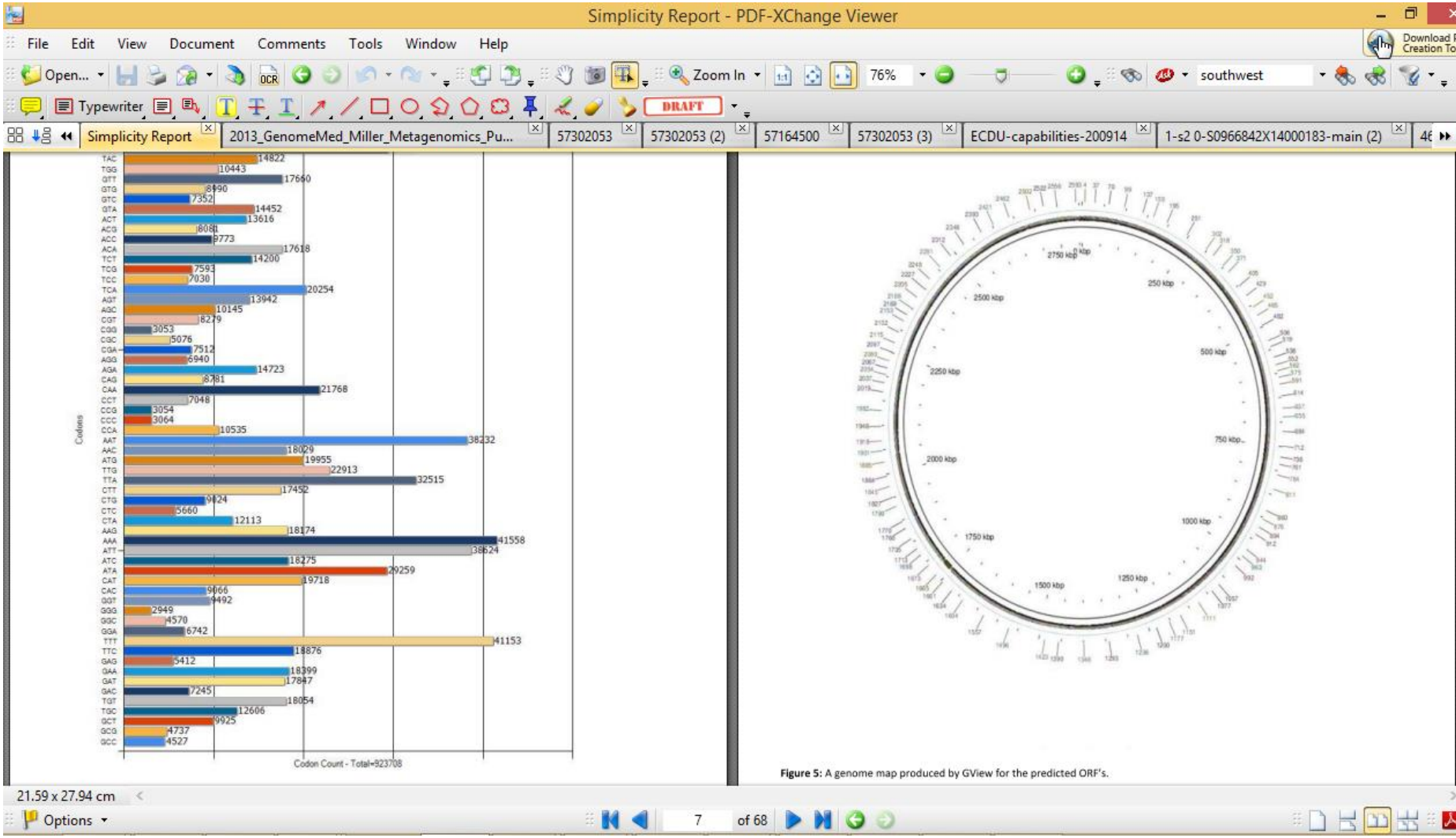
Per base sequence content	FAIL	FAIL
Per base GC content	WARN	WARN
Per sequence GC content	PASS	PASS
Per base N content	PASS	PASS
Sequence Length Distribution	WARN	WARN
Sequence Duplication Levels	WARN	WARN
Overrepresented sequences	PASS	PASS
Kmer Content	PASS	WARN

Table 2: High level overview of tests performed by FastQC.

21.59 x 27.94 cm

Options 1 of 68

Reports


Simplicity Report - PDF-XChange Viewer
File Edit View Document Comments Tools Window Help
Simplicity Report
2013_GenomeMed_Miller_Metagenomics_Pu...
57302053
57302053 (2)
57164500
57302053 (3)
ECDU-capabilities-200914
1-s2 0-S0966842X14000183-main (2)
Codons
Codon Count - Total=923708
Figure 5: A genome map produced by GView for the predicted ORF's.
21.59 x 27.94 cm
Options
7
of 68

Reports

Simplicity Report - PDF-XChange Viewer

File Edit View Document Comments Tools Window Help

Simplicity Report | 2013_GenomeMed_Miller_Metagenomics_Pu... | 57302053 | 57302053 (2) | 57164500 | 57302053 (3) | ECDU-capabilities-200914 | 1-s2 0-S0966842X14000183-main (2)

betaine-aldehyde dehydrogenase activity	1
beta-ketoacyl-acyl-carrier-protein synthase III activity	1
aspartate 1-decarboxylase activity	1
aspartate carbamoyltransferase activity	1
aspartate kinase activity	1
aspartate-semialdehyde dehydrogenase activity	1
aspartate-tRNA ligase activity	1
antiporter activity	1
ATP citrate synthase activity	1
ATP phosphoribosyltransferase activity	1
arginine deiminase activity	1
arginine:ornithine antiporter activity	1
arginine-tRNA ligase activity	1
arsenate reductase (thioredoxin) activity	1
arsenite transmembrane transporter activity	1
ammonia-lyase activity	1
ammonium transmembrane transporter activity	1
AMP binding	1
anthranilate phosphoribosyltransferase activity	1

Page N°29

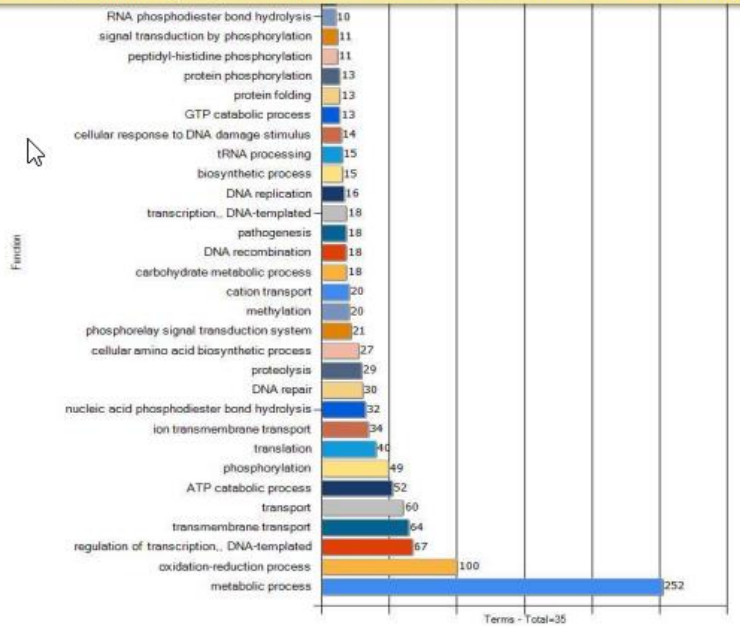


Figure 8: Gene Ontology terms for biological function associated with the top hits.

Page N°30

21.59 x 27.94 cm

Options 29 of 68

Health Record Management

Simplicity _EHR

- Customizable cloud- based platform
- Designed by clinicians for clinicians
- Heterogeneous data – clinical terms – SNOMED, OpenEHR and UMLS
- Melanoma and Colorectal cancer patients
- Built to FDA 21CFR11 guidelines.
- Uses HIPPA mandated .net infrastructure.
- ISO/TR 20514 - Health informatics — Electronic health record — Definition, scope and context
- See demo….
  http://youtu.be/JBxjSY67iMw

Joe Byrne

Edit | No need for further discussion

MRN	243234ABC	**Initial diagnosis**	Lower limb
Age	17 years old	**Number of MDTs**	4
Date of birth	09/06/1997	**Last MDT**	07/07/2014
Address	Cork	**Next meeting**	23/07/2014

Lesions

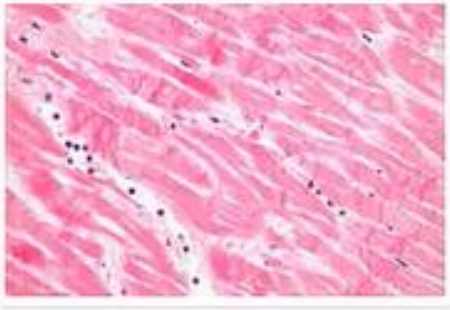
- Lower limb
- Trunk

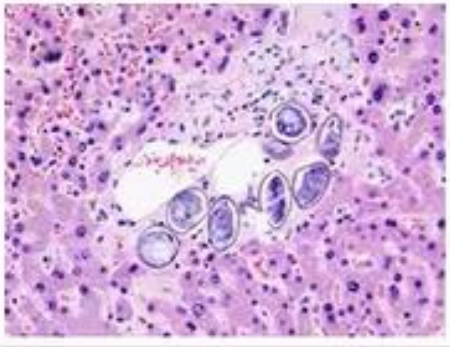
Site of lesion	Lower limb
Specific site	Toe
Breslow thickness	1 mm
Mitosis	8 mm²
Ulceration	False
T-Stage	pT3a

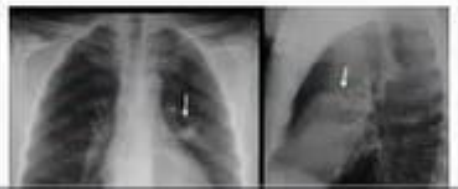
Date 11/06/2014

☐ Clinical Details Available ☑ Photo Available ☐ Slides available for review

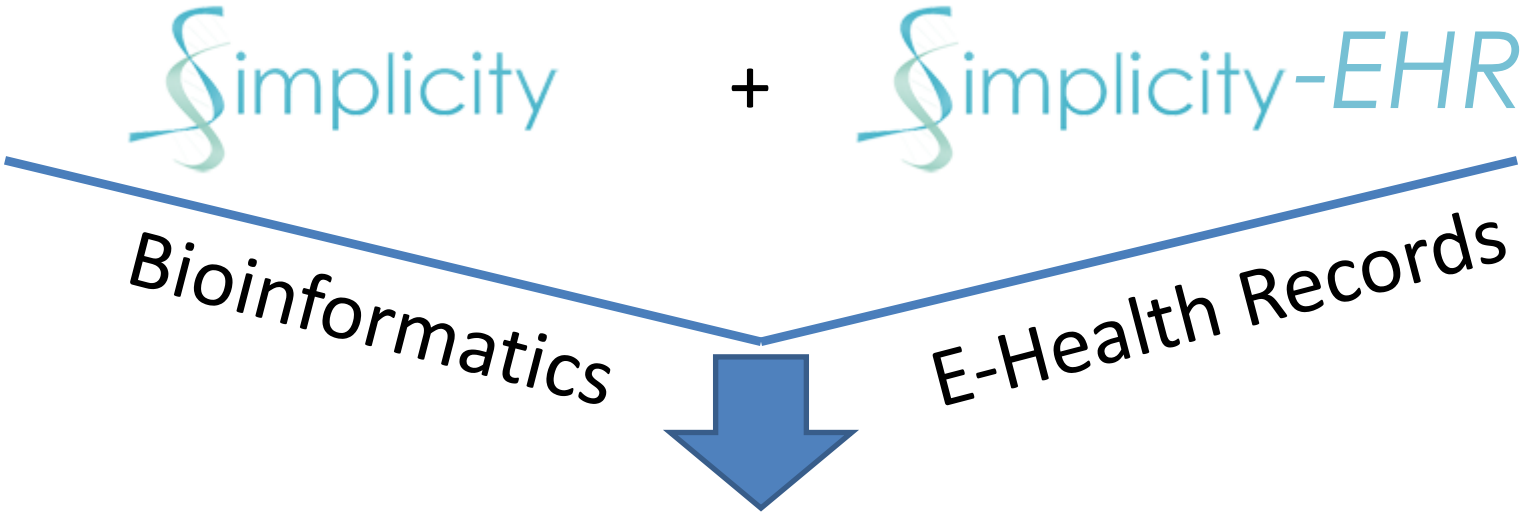
Upload pictures







- Confluence of our two product areas



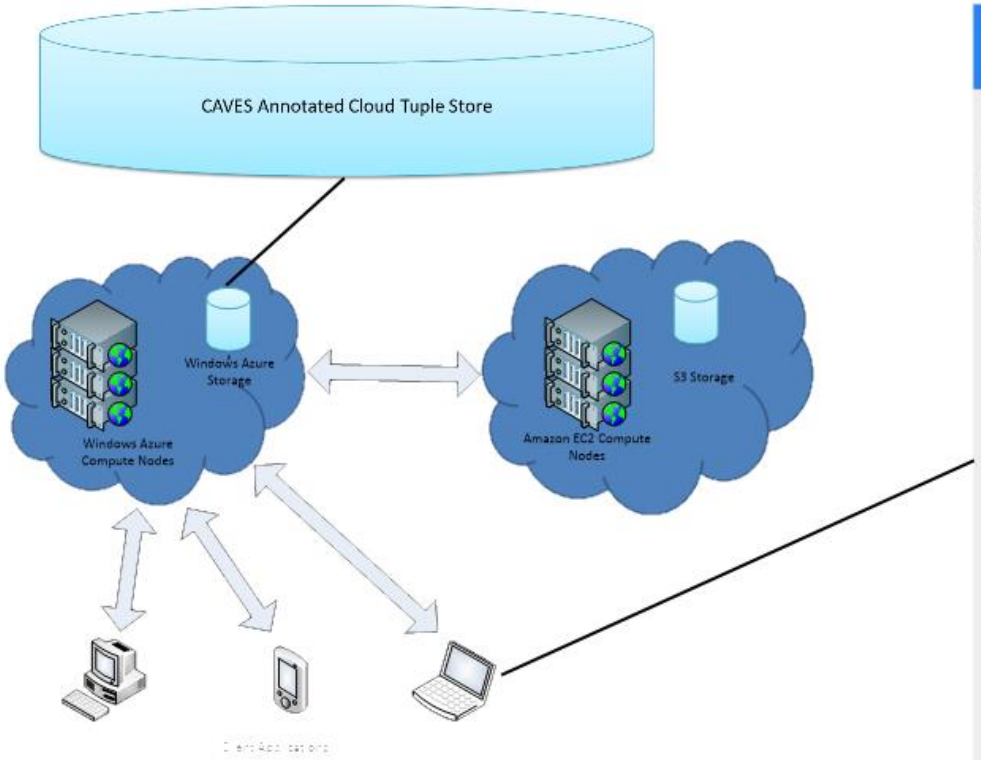
Proprietary tools which accelerate the development of:

- Molecular Diagnostics
- Clinical Care
- Personalised Medicine

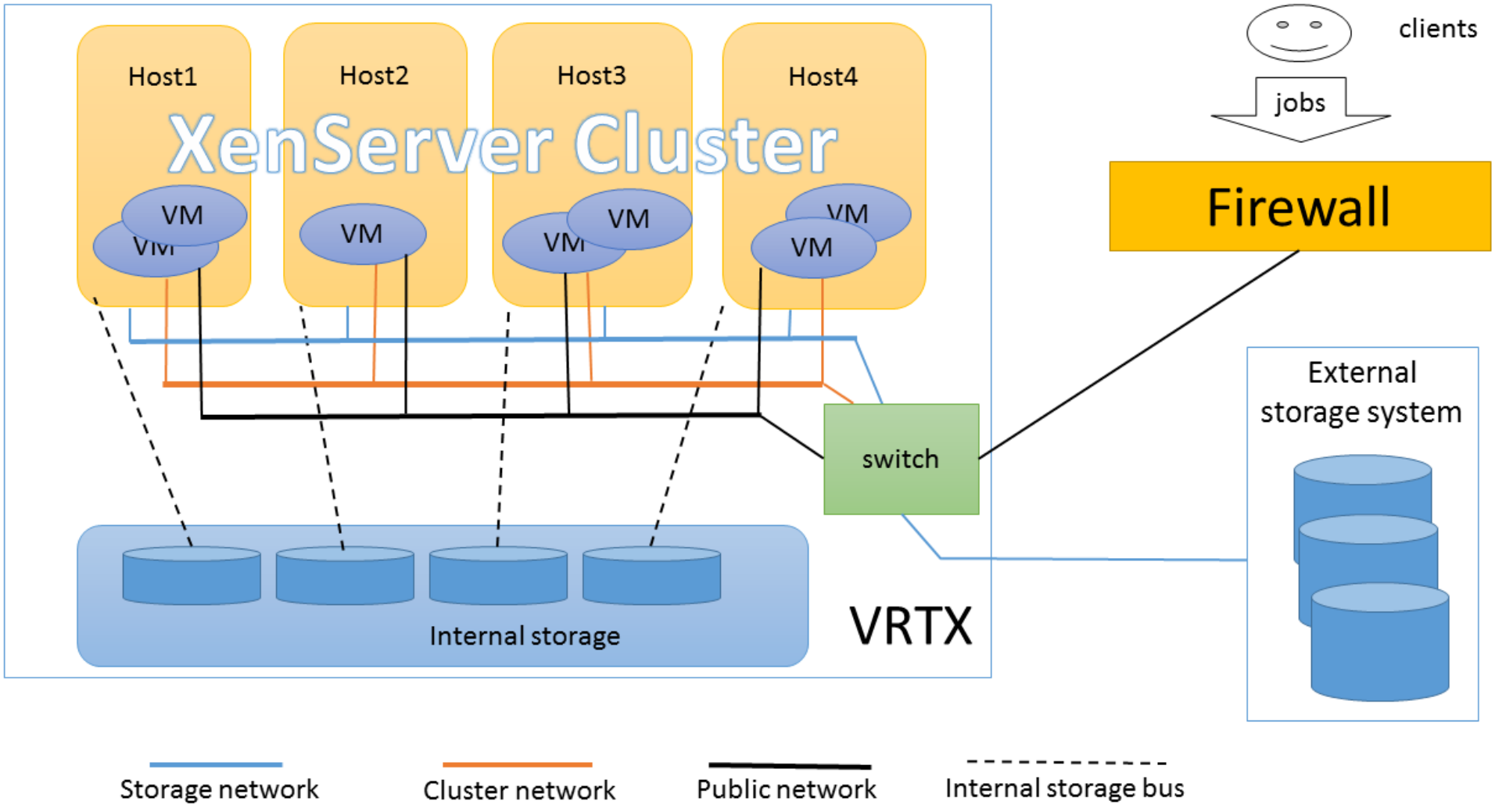
System Architecture

- Configuration and Traceability are all enabled by mixed schemas.
- Relational databases are used for ‘classic’ data management via SQL – column modelling
- Sparse medical data treated as ‘row modelled’ entities with attribute meta data – object store.
- XML used for open schemas such as user sessions, new tools and OpenEHRs
- JSON used to describe services.
- BLOBS – Binary Large Objects used for multimedia object storage.

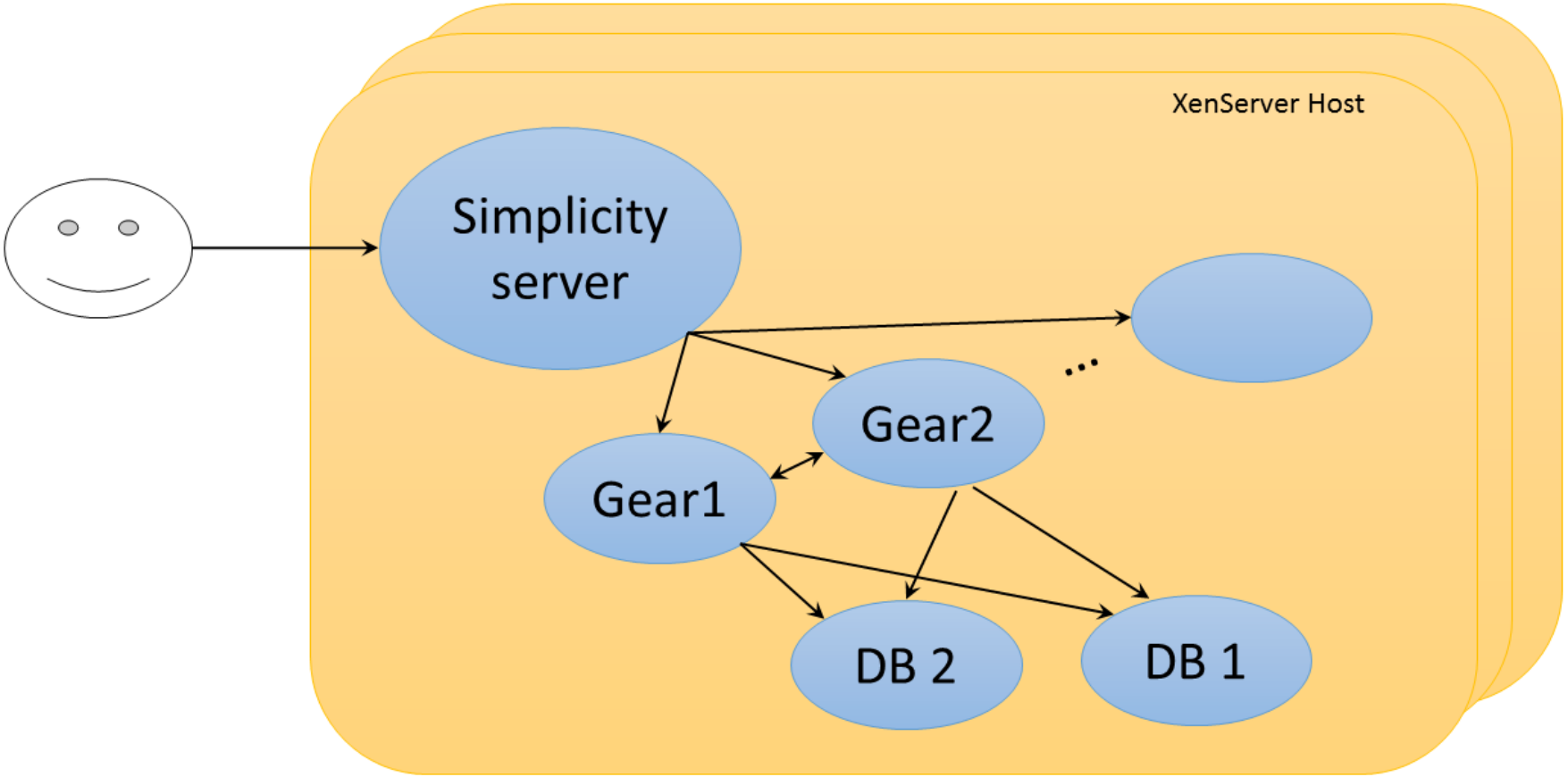
System Architecture



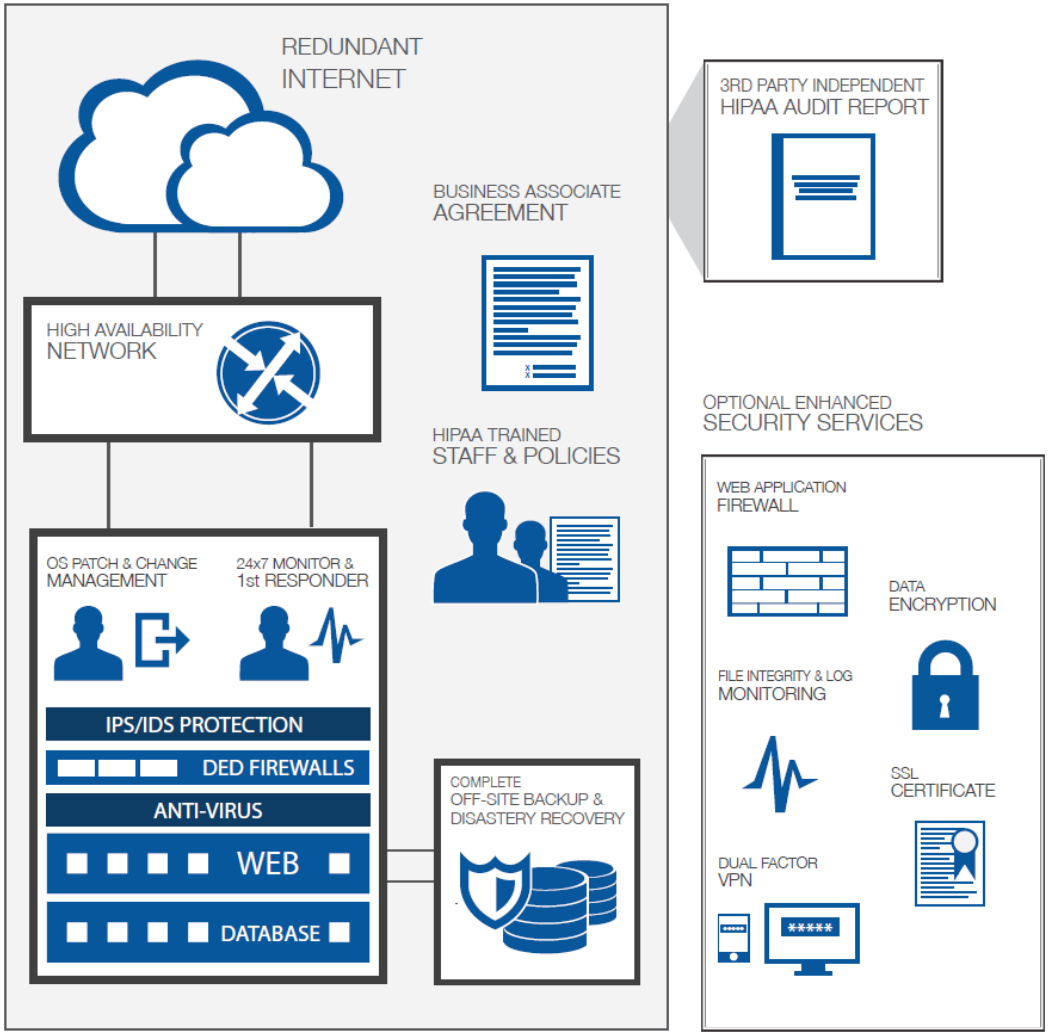
System Architecture – Virtual Nodes



Virtual Machines Roles



Security HIPAA

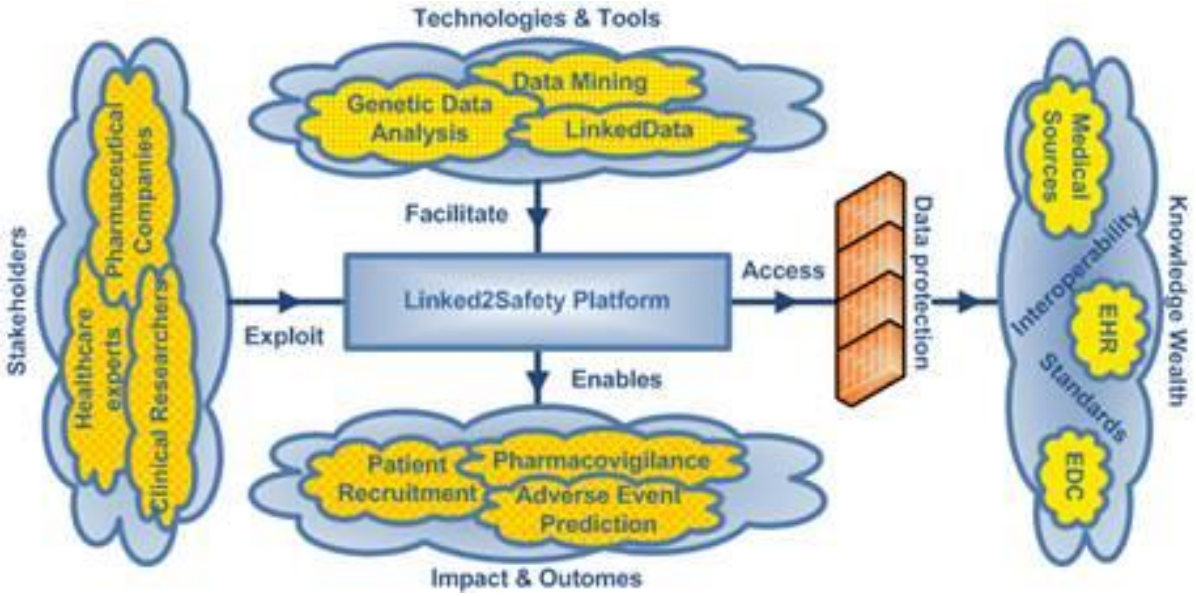


FDA 21 CFR 11

- Limiting system access to authorized individuals
- Use of secure, computer generated, time-stamped audit trails.
- Use of authority checks to ensure that only authorized individuals can use the system, electronically sign a record and so on…..

Patient Privacy

- **Linked2Safety Platform - dynamically interconnecting distributed patients data with clinical research efforts**, **respecting patients' anonymity**, data ownership and privacy, as well as **European and national legislation**.



Archiving

- Best practice: 3 copies, 3 locations, online and offline

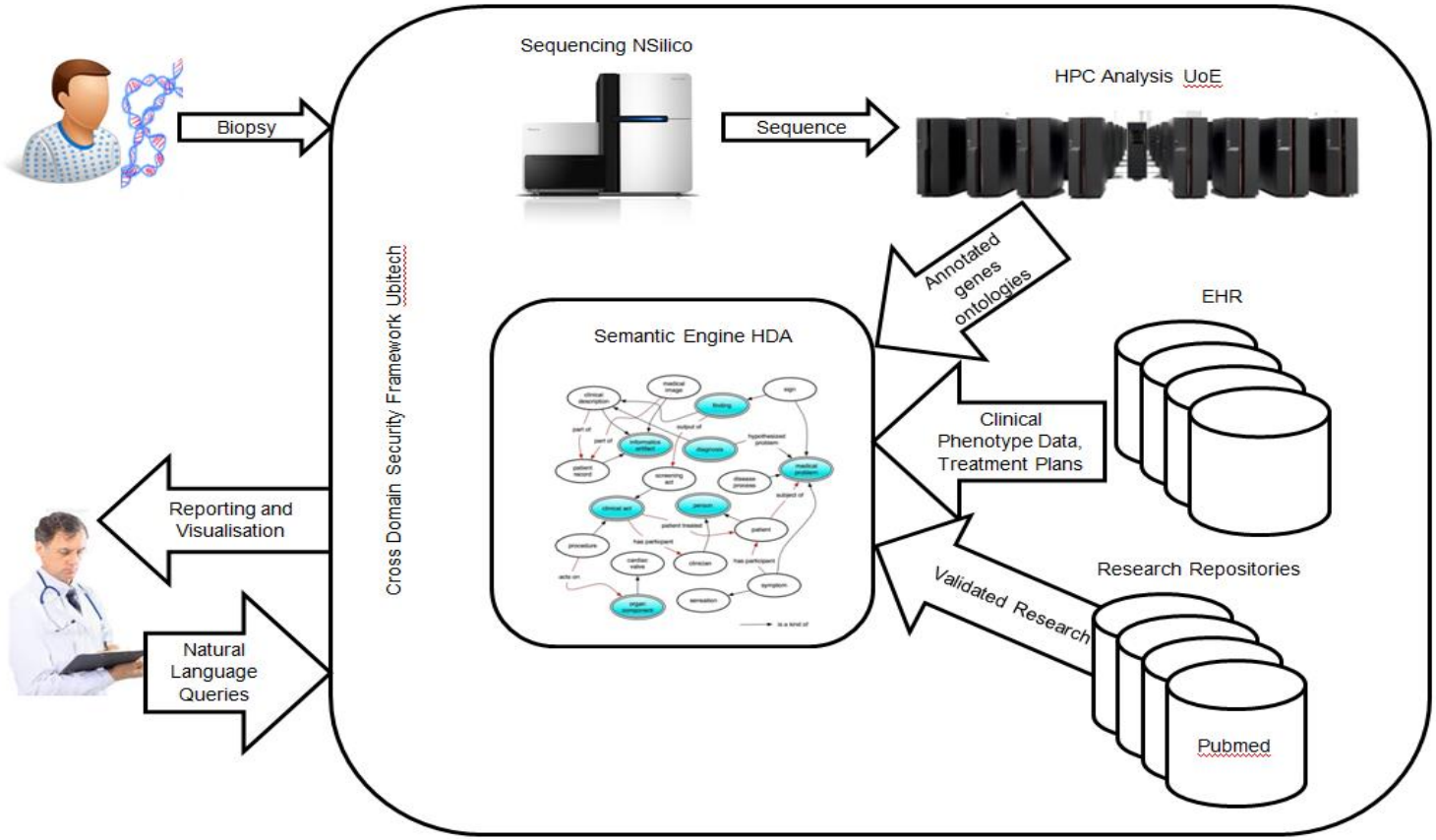
- Active archiving: integrity, obsolescence

- Risk management: ISO27001

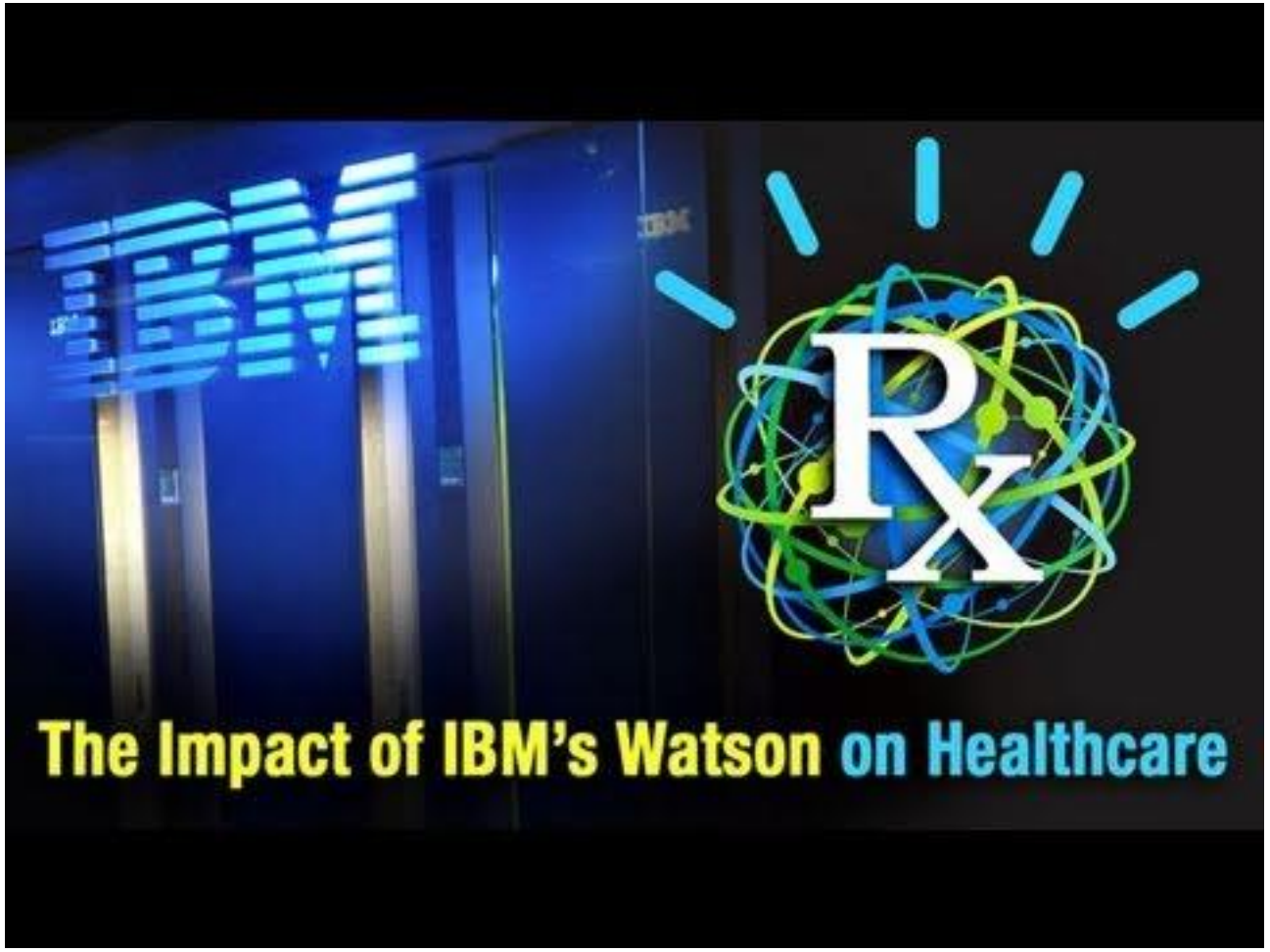
Firewalled Compute Appliance

Next Steps

- Applying same approach to melanoma.
- Semantically integrate unstructured data.
- Just recently won H2020 to tackle this…SAGE- CARE



IBM Partnership



Partners

- Currently collaborating to further develop our platforms and products which leverage big data for clinical research

FDA

Microsoft

DELL

IBM

Health Innovation Hub

Division of PATHWAY MEDICINE

THE UNIVERSITY OF EDINBURGH

ICHEC
Irish Centre for High-End Computing

CIT
Institiúid Teicneolaíochta Chorcaí
Cork Institute of Technology

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Conclusions

- Wide range of issues to be addressed.
- Usability, Security, Privacy, Responsiveness, Size, Accuracy
- Solutions
  - **Speed** – Cloud, Multicore, Muli-threaded
  - **Scalability** – Cloud Elasticity, storage, S3 , AWS Glacier,
  - **Security** - HIPPA, 21 CFR 11, L2S, Firewall, Traceability
  - **Simplicity** - Usability

Thank You

Thanks to Prof Peter Ghazal, Clair Smith, Kate Templeton, Paul Dickinson, Kai Kropp, Vladimir Belogrudov, Aisling O'Driscoll, Roy Sleator, Brian Kelly
DPM, CIT, REI, NSilico

Paul.walsh@nsilico.com